ELSEVIER

Available online at www.sciencedirect.com

ScienceDirect

Speech Communication

SPEECH COMMUNICATION

www.elsevier.com/locate/specom

Special Issue: Phase-Aware Signal Processing in Speech Communication

Synthetic Speech Detection Using Phase Information

Ibon Saratxaga[a], Jon Sanchez*[a], Zhizheng Wu[b], Inma Hernaez[a]

[a]*University of the Basque Country, UPV- EHU, Spain*
[b]*The Centre for Speech Technology Research (CSTR), University of Edinburgh, UK*

Abstract

Taking advantage of the fact that most of the speech processing techniques neglect the phase information, we seek to detect phase perturbations in order to prevent synthetic impostors attacking Speaker Verification systems. Two Synthetic Speech Detection (SSD) systems that use spectral phase related information are reviewed and evaluated in this work: one based on the Modified Group Delay (MGD), and the other based on the Relative Phase Shift, (RPS). A classical module-based MFCC system is also used as baseline. Different training strategies are proposed and evaluated using both real spoofing samples and copy-synthesized signals from the natural ones, aiming to alleviate the issue of getting real data to train the systems. The recently published ASVSpoof2015 database is used for training and evaluation. Performance with completely unrelated data is also checked using synthetic speech from the Blizzard Challenge as evaluation material. The results prove that phase information can be successfully used for the SSD task even with unknown attacks.

Keywords: Synthetic speech detection, Phase, MGD, RPS

* Corresponding author. Tel.: +34-946014122; fax: +34-946014259.
E-mail address: ion@aholab.ehu.eus

1. Introduction

In speech processing, in speech synthesis and analysis areas alike, phase information has been traditionally discarded for most of the conventional applications. The spectral module information is highly correlated with the perceptual features of the speech and there are well established techniques to process them. Phase information has subtler perceptual effects (Alsteris and Paliwal, 2007) (Saratxaga et al., 2012) and tricky features like wrapping make it more difficult to model and process.

This unawareness for phase information in most speech processing techniques can indeed be exploited to detect such a processing on speech, tracing the unintended perturbations of the natural phase patterns left behind by this processing. One particular case where detecting natural speech manipulations can be critical is the speaker verification field.

The first speaker verification (SV) systems tried to resolve the problem of detecting if a voice was certainly from a claimant speaker and not from other (Rosenberg, 1976). The improvement of the SV systems allowed a high success rate solving the problem of naive speaker verification, but the parallel advance in speech manipulation techniques has posed a new menace to these systems: impostors forging speech signals that imitate a particular speaker's voice. This threat was first pointed by (Pellom and Hansen, 1999) and (Masuko et al., 2000), and has received more and more attention in literature as new voice adaptation and transformation techniques have made more feasible to mimic a speaker's voice with less and less material from the original speaker. A detailed survey is published in (Wu et al., 2015).

Nowadays two are the main speech processing techniques that allow the creation of synthetic speech spoofing signals: First, the statistical speech synthesizers (Yoshimura et al., 1999) (Tokuda et al., 2002) using voices adapted to a particular speaker (Yamagishi et al., 2009) even with minimum quality material (Yamagishi et al., 2010). Second, the voice conversion (VC) techniques (Jin et al., 2008), (Kinnunen et al., 2012). Both techniques can be used to generate spoofing signals that can successfully deceive state-of-the-art SV systems with false acceptance rates (FAR) around 80% for synthetic speech and 5% for VC.

A number of countermeasures have been proposed to these attacks. In (Satoh et al., 2001), a countermeasure based on the average inter-frame difference was proposed to discriminate between natural and synthetic speech from an HMM-based speech synthesis system. Another similar countermeasure which also use an average pair-wise distance between consecutive frames was proposed to detect voice-converted speech (Alegre et al., 2013a). Rather than capturing the inter-frame distortions, in (Wu et al., 2013) and (Alegre et al., 2013b), modulation-based features and local binary pattern-based features were proposed to utilize long-term spectro-temporal information for synthetic speech detection. In (Sizov et al., 2015), a countermeasure which uses the same front-end as ASV was proposed to discriminate natural and voice-converted speech.

Above countermeasures which derive features from magnitude spectra work well for the specific attacks. Phase-based countermeasures proposed by the authors of this work have been used for both synthetic and voice-converted speech detection. In (Wu et al., 2012) synthetic speech detectors (SSD) based on cosine normalized phase and modified-group delay (MGD) (Yegnanarayana and Murthy, 1992) are evaluated with converted spoofing signals. In (Wu et al., 2013), modulation spectrum derived from the modified group delay spectrum was used for synthetic speech detection. These works have confirmed the effectiveness of phase information in detecting synthetic speech with matched vocoder.

Relative Phase Shift (RPS) representation (Saratxaga et al., 2009) for the harmonic phase has also be used to build SSD systems aimed to detect spoofing signals created with adapted synthetic voices (De Leon et al., 2011) (De Leon et al., 2012) with good results. The initial works were focused on evaluating the actual capability of the RPSs to detect the phase modifications due to the synthetic generation of the spoofing signals. Consequently synthesized impostors were used to model the spoofing attacks. This approach has the double downside of requiring the adaptation of synthetic voices to generate the spoofing samples, and, more important, using particular attacks to train the synthetic models yields that their performance will be attack-dependent, and they will not be able to detect spoofing signals created with another attacking technique.

Once the validity of the RPS based SSD was demonstrated, the problem of avoiding attack dependence of the SSD was addressed in (Sanchez et al., 2014) (Sanchez et al., 2015). In these works, the authors analyze the use of copy-synthesized signals to create the imposter models. This way, the models are not dependent on the particular features of a specific synthesizer, but they can detect any signal created with a vocoder. Multi-vocoder models trained and tested with completely unrelated signals were evaluated with good results.

Recently the work in this area has been promoted by the ASVSpoof2015, the Automatic Speaker Verification Spoofing and Countermeasures Challenge (Wu et al., 2014). The participants were invited to submit the results of independent SSD modules for evaluation. Spoofing detection systems were tested with a database (the so-called ASVSpoof database), containing different spoofing techniques such as speech synthesis and voice conversion. The performance of the different systems was assessed by the organization using standard metrics. This database has been made available to the public, and we are using it in this work.

In this paper we review and evaluate the performance of a MGD based and a RPS based SSD system, benchmarking them against a module information based (MFCC) baseline system. We analyze the optimal use of training material comparing the strategy of using “real” spoofing signals versus using copy-synthesis signals from the natural ones. The performance of the system with completely unknown signals is also evaluated using a completely unrelated set of signals from the Blizzard Challenge (Black and Tokuda, 2005), the most popular international event for TTS system evaluations, where independent participants build synthetic voices using a common speech corpus and send some samples to be evaluated. They are, undoubtedly, a representative sample of the current technology in speech synthesis, and, consequently, of the kind of likely spoofing technique.

Furthermore, the tests with a completely unrelated database, as the Blizzard Challenge one, introduces the channel-mismatch issue for spoofing detection. While in the ASVSpoof Challenge the same recording channel is assumed for every signal, the channel information of Blizzard Challenge data is different from ASVSpoof data. The robustness to the channel of the different SSDs has been little studied in literature and will be analyzed in this work for the proposed systems.

The paper is organized as follows. First, the phase representation and parameterization methods – RPS and MGD – are described. Then, in section 3, the Synthetic Speech Detection System is described. 4th section is devoted to describe the databases used in both the training and test phases, and in the 5th section the evaluation experiments are detailed. Finally, some conclusions are drawn.

2. Phase representation and parameterization.

We will evaluate two different phase-based systems: the Relative Phase Shift (RPS), based on the phase shift of the harmonic components of the speech signal, and the Modified Group Delay (MGD), which includes both magnitude and phase related information. Both systems are described below.

2.1. Relative Phase Shift (RPS)

The Relative Phase Shift (RPS) is a representation for the phase information of a harmonic speech signal. The representation was derived in (Saratxaga et al., 2009), but a brief description is provided in this section.

2.1.1. Definition and derivation

RPS is a representation for the harmonic phase. Harmonic analysis models each frame of a signal by means of a sum of sinusoids harmonically related to the pitch or fundamental frequency.

$$h(t)=\sum_{k=1}^{N} A_k \cos\left(\varphi_k\left(t\right)\right) \tag{1}$$

$$\varphi_k\left(t\right)=2\pi k f_o t+\theta_k \tag{2}$$

where N is the number of bands, A_k the amplitudes, $\varphi_k(t)$ the instantaneous phases, f_0 the pitch or fundamental frequency and θ_k is the initial phase shift of the k-th sinusoid. The instantaneous phase is composed of two terms: the so-called "linear component" (depending on the analysis time instant and the frequency of the harmonic) and the initial phase shift term. This complex dependency makes the instantaneous phase difficult to use for certain purposes (most notably for pattern analysis and statistical modeling).

The RPS representation consists in calculating the phase shift between every harmonic and the fundamental component (k=1) at a specific point of the fundamental period, namely the point where φ_1=0. This does not imply that the analysis has to be done at that specific time (i.e. pitch synchronous), by the contrary, assuming local stationarity, the RPS value can be calculated at any time analysis instant. Let us consider two sinusoidal harmonic components like:

$$\begin{aligned} x_1(t) &= \cos\left(2\pi f_0 t+\theta_1\right) \\ x_k(t) &= \cos\left(2\pi k f_0 t+\theta_k\right) \end{aligned} \tag{3}$$

In a analysis time instant t_a the instantaneous phase of each component will be:

$$\begin{aligned} \phi_1(t_a) &= 2\pi f_0 t_a+\theta_1 \\ \phi_k(t_a) &= 2\pi k f_0 t_a+\theta_k \end{aligned} \tag{4}$$

For the RPS ψ we have to calculate the phase shift in the instant t_o the closest instant before the analysis point when $\varphi_1(t_o)$=0, as

$$\psi(t_a)=\phi_k(t_o)-\phi_1(t_o)=\phi_k(t_o) \tag{5}$$

Assuming local stationarity, we can extrapolate the value of the instantaneous phase of the k-th harmonic. If we use principal values for the phases for simplicity, $\varphi_1(t_o)$=0 and we can obtain t_o from (4):

$$t_o=\frac{-\theta_1}{2\pi f_0} \tag{6}$$

From (4) we also know that:

$$\theta_1=\varphi_1\left(t_a\right)-2\pi f_0 t_a \tag{7}$$

Combining (4), (6) and (7) in (5), we have:

$$\psi(t_a) = \varphi_k(t_o) = 2\pi k f_0 \left(t_a - \frac{\varphi_1(t_a)}{2\pi f_0} \right) + \theta_k = 2\pi k f_0 t_a + \theta_k - k\varphi_1(t_a) \tag{8}$$

And so we obtain the RPS transformation for the *k*-th harmonic component, whose graphical interpretation is shown in Figure 1:

$$\psi_k(t_a) = \varphi_k(t_a) - k\varphi_1(t_a) \tag{9}$$

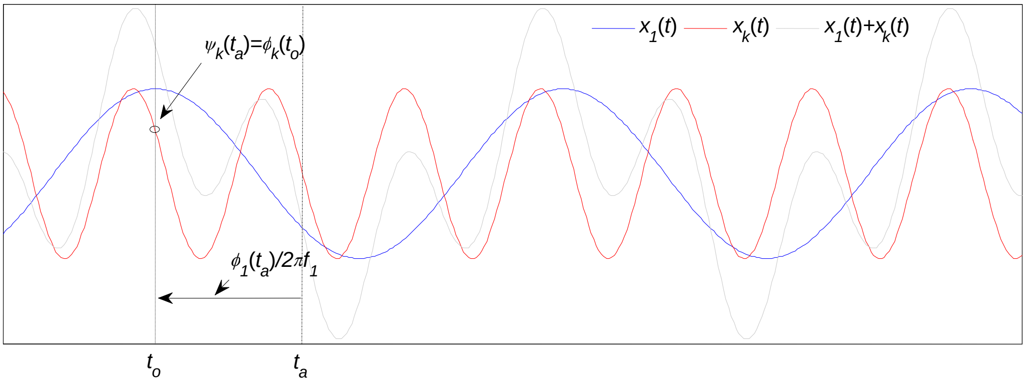

Figure 1: *Graphical interpretation of the RPS transformation: for an analysis instant t_a the RPS of k is the phase shift of that component with respect to the fundamental component at the point where the period of the fundamental component starts (t_o).*

Equation (9) defines the RPS transformation which allows computing the RPSs (ψ_k) from the instantaneous phases at any point (t_a) of the signal. The RPS values are wrapped to the [-π, π] interval.

The RPS transformation intrinsically removes the linear phase term, thus resulting in a magnitude that remains stable as long as the phase shift relations of the components (and subsequently the waveform) do not change. These stable patterns allow the phase structure to arise, as is shown in Fig. 2 where instantaneous phase (a) and RPS values (b) of a voiced speech signal /aeiou/ (c) can be compared. It is worth noting that there is no useful phase information in the unvoiced frames of the speech.

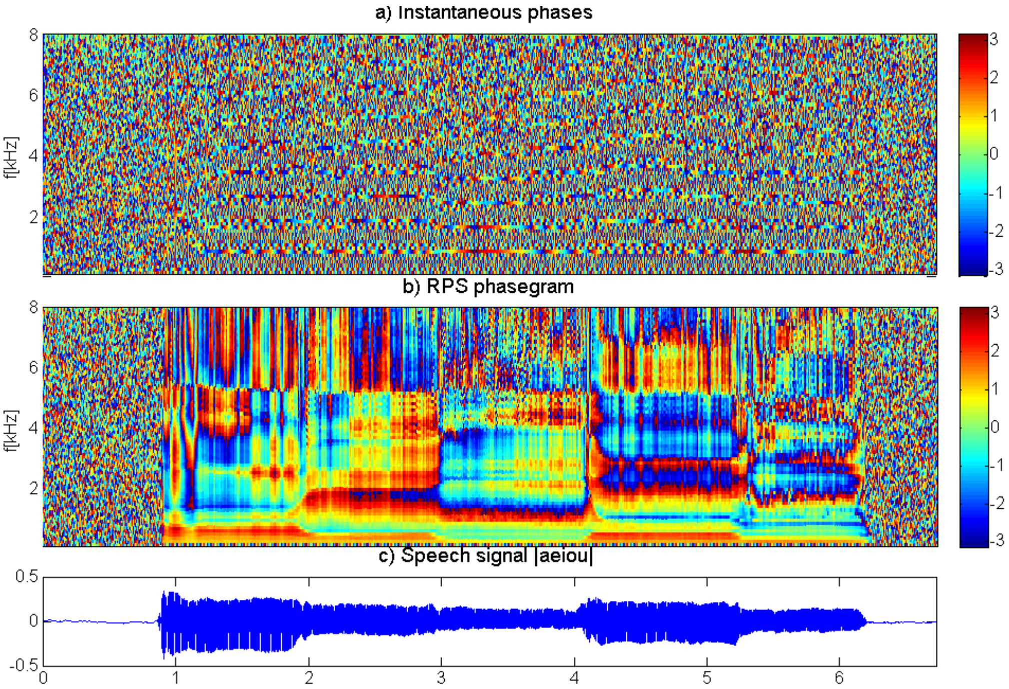

Figure 2: *Phasegrams of a voiced speech segment with five continuous vowels. (a) Instantaneous phases. (b)Relative phase shift (c) Signal waveform.*

2.1.2. Parameterization

Although the RPS patterns look very definite, the RPS values are not suitable for statistical modeling. Variable numbers of values depending on the number of harmonics, high dimensionality, wrapping discontinuities, etc. make it necessary to apply additional parameterization.

In (Saratxaga et al., 2010) it was described a method to obtain a reduced parameter set out of the RPS values, the so called DCT-mel-RPS parameterization. This parameterization reduces the variable number of raw RPS values to a constant number of parameters and is well suited for statistical modeling.

To obtain the parameters, the differences of the unwrapped RPS values are filtered with a mel filter bank (48 filters) and a discrete cosine transform (DCT) is applied to the resulting sequence. The DCT is truncated to 20 values and the averaged value of the slope of the unwrapped RPS values is also included. The Δ and ΔΔ values of this vector are calculated which leads to a total of 63 phase-based parameters, calculated only for voiced frames, usually with frame rates of 5-10ms.

2.2. MGD

The modified group delay (MGD) feature is a representation of complex Fourier transform spectrum, and contains both magnitude and phase spectra information. It has been used for speech recognition in (Zhu and Paliwal, 2004) and (Hegde et al., 2007). This section briefly introduces MGD feature.

Given a speech signal $x(n)$, the complex spectrum representation $X(\omega)$ can be obtained through short-time Fourier transform. The complex spectrum $X(\omega)$ has two parts: real part $X_R(\omega)$ and imaginary part $X_I(\omega)$. The power spectrum which derives the popular Mel-Frequency Cepstral Coefficients (MFCC) is represented as

$|X(\omega)|^2$. To extract modified group delay spectrum, we define $Y(\omega)$ as the complex spectrum of $nx(n)$, which is a re-scaled signal of $x(n)$. The modified group delay spectrum $\tau_{\rho,\gamma}(\omega)$ is defined as,

$$\tau_{\rho}(\omega)=\frac{X_R(\omega)Y_R(\omega)+Y_I(\omega)X_I(\omega)}{|S(\omega)|^{2\rho}} \tag{10}$$

$$\tau_{\rho,\gamma}(\omega)=\frac{\tau_{\rho}(\omega)}{|\tau_{\rho}(\omega)|}|\tau_{\rho}(\omega)|^{\gamma} \tag{11}$$

where $X_R(\omega)$ and $X_I(\omega)$ are the real and imaginary parts of $X(\omega)$, respectively, $Y_R(\omega)$ and $Y_I(\omega)$ are the real and imaginary parts of $Y(\omega)$, $|S(\omega)|^2$ is the smoothed power spectrum corresponding to $|X(\omega)|^2$, and ρ and γ are two weighted variables to control the shape of the modified group delay spectrum. In practice, $|S(\omega)|^2$ is obtained by cepstrally smoothing the power spectrum $|X(\omega)|^2$. This can be achieved through two steps:

a) apply discrete cosine transform (DCT) on the power spectrum and

b) then pass the first 30 DCT coefficients to inverse discrete cosine transform (IDCT) to reconstruct a new smoothed spectrum.

The reason to use the smoothed spectrum rather than the original spectrum is to make the modified group delay spectrum much more stable (Hegde et al., 2007). A spectrogram-like graphical representation of this magnitude is shown in Figure 3.

With the modified group delay spectrum, we can compute modified group delay cepstral coefficients (MGDCC) as feature representations for modeling. The cepstral feature can be computed through the following steps:

a) Apply Fourier transform to the signal $x(n)$ and its re-scaled version $nx(n)$ to compute the spectrum $X(\omega)$ and $Y(\omega)$, respectively.

b) Compute the cepstrally smoothed spectrum $|S(\omega)|^2$ for the power spectrum $|X(\omega)|^2$.

c) Compute modified group delay spectrum using Equation (10) and (11).

e) Apply DCT on the modified group delay spectrum to calculate the MGDCC.

The two controlling variable ρ and γ are tuned on the development set for better representation performance.

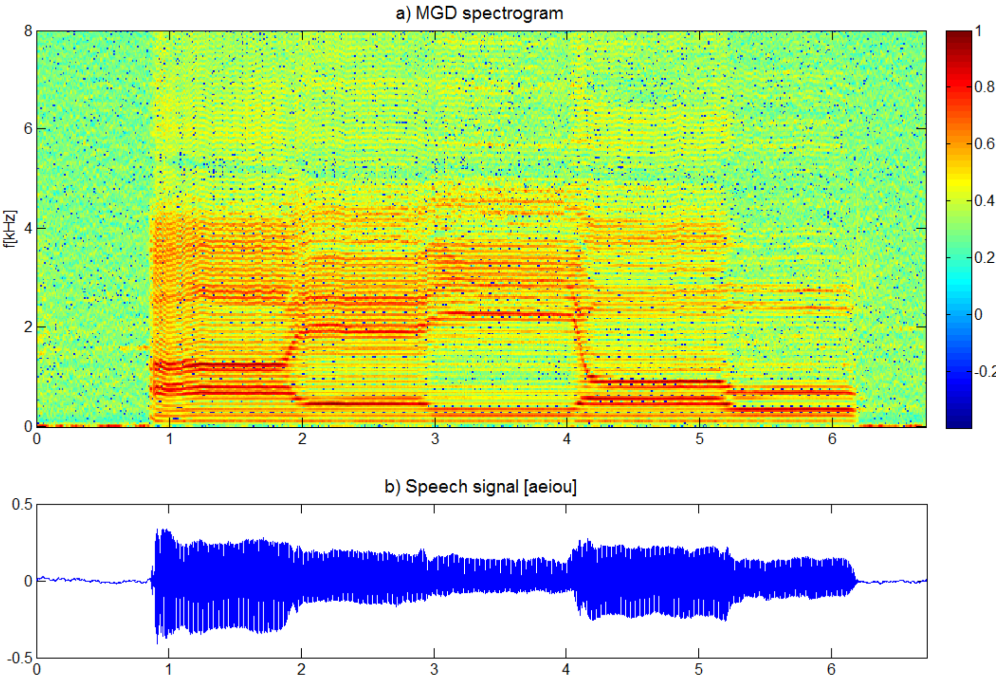

Figure 3: *MGD spectrogram of a voiced speech segment with five continuous vowels.*

3. Synthetic Speech Detectors (SSD)

In this work we will compare different Synthetic Speech Detectors (SSD) systems. The purpose of the SSD systems is to discriminate between natural speech signals and synthetically generated ones. SSD blocks are intended to work jointly with speaker verification (SV) systems, trying to detect synthetically generated speaker adapted impostor signals that can cheat the SV system. If the SSD system requires knowing the supposed speaker identity to perform the classification task (i.e. it uses speaker dependent models) then the SSD will necessarily be placed after the SV system to check the signals accepted as claimants by the SV system. If previous knowledge of the speaker identity is not necessary (i.e. speaker independent models), the SSD module can be inserted before or after the SV system. This is the case of the systems analyzed in this work.

Figure 4 shows the main structure of an SSD system. The system is a binary classifier. During the training phase, parametric models for both natural speech (λ_{human}) and synthetic speech (λ_{synth}) are created. Then, candidate parameter vectors are evaluated.

To perform the synthetic speech detection task, the system will test a candidate vector sequence $Y=\{\mathbf{y}_1,\ldots,\mathbf{y}_N\}$ of length N against both natural speech and synthetic speech models to get the corresponding likelihood values $p(Y|\lambda_{\text{human}})$ and $p(Y|\lambda_{\text{synth}})$. Then, the log likelihood ratio Λ is calculated as

$$\Lambda(Y) = \log p(Y|\lambda_{\text{human}}) - \log p(Y|\lambda_{\text{synth}}) \qquad (12)$$

where

$$\log p(Y|\lambda) = \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{y}_n|\lambda) \qquad (13)$$

The candidate is considered human if it exceeds a certain decision threshold θ which was set to the equal error rate (EER) point in the experiments.

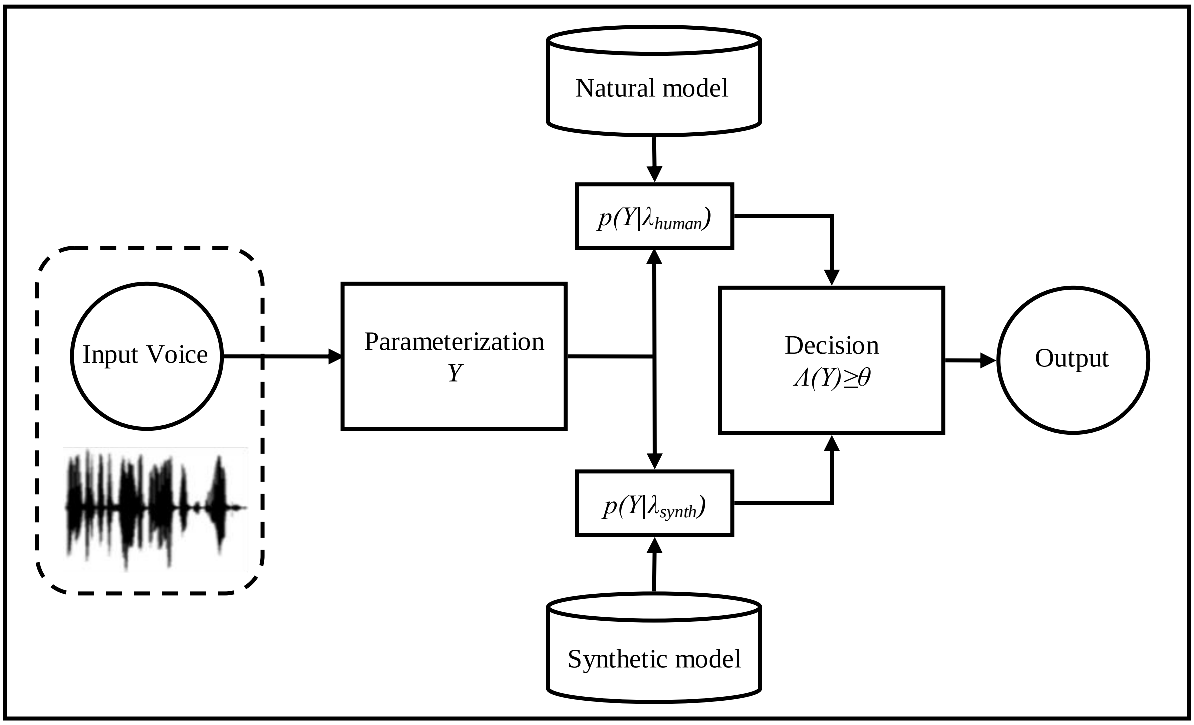

Figure 4: *SSD system structure.*

In this paper three different systems are referred, with different parameterization and modelling techniques. The first one, MFCC, based on the spectral module information, is included as a baseline. The second one, MGD has been successfully used for SSD experiments (Wu et al., 2013). The third one, RPS, has been also previously tested in different spoofing scenarios (Sanchez et al., 2014) (Sanchez et al., 2015). In this paper, both phase-based systems will be facing new spoofing experiments and compared.

3.1. Natural and synthetic models

In this work, we focus on feature-based countermeasures rather than model-based approaches. Hence, we use the classic GMM-based classifier for the detectors. The GMM-based classifiers have 1024 Gaussian components for the MFCC and MGD models and 2048 components for the RPS based models. The natural models are trained on the training data of human speech defined by ASVSpoof 2015 protocol, while the synthetic models are trained on the training data of the five known attacks (also as defined in ASVSpoof 2015 protocol) , and/or copy-synthesis speech as it will be described in section 5.

4. Training and evaluation databases

4.1. ASVSpoof Database

This database was created for the Automatic Speaker Verification Spoofing and Countermeasures Challenge (Wu et al., 2014).

The natural speech information was collected from 106 speakers (61 female and 45 male). There are no remarkable channel or background noise effects. Taking these genuine human signals as a basis, different spoofing algorithms are selected to create the spoofed speech. The signals are originally sampled at 16 kHz, and that is how they are used to calculate MFCC and MGD parameters, for DCT-mel-RPS they have been downsampled to 8 kHz before being parameterized.

In order to perform training, evaluation and testing, the whole data base is divided in three datasets. Different speakers are selected for each of the sets. The number of speakers and in each dataset is illustrated in Table 1.

Table 1: Number speakers and utterances in the different datasets (Wu et al., 2015b)

	#Speakers		#Utterances	
Subset	Male	Female	Genuine	Spoofed
Training	10	15	3750	12625
Development	15	20	3497	49875
Evaluation	20	26	9404	184000

4.1.1. Training data

25 speakers, 15 female and 10 male, were selected to make up the training data set. Together with the genuine voice utterances, the spoofed versions are also part of the set, created by means of five different systems: three of them voice conversion based (including frame selection and spectral slope techniques, and a publicly available voice conversion toolkit within the Festvox system[†]) and two speech synthesis algorithms (both implemented using HMMs and HTS[‡]).

4.1.2. Development data

The second subset of the database, intended to be used for development, takes 3497 genuine utterances from 35 speakers (20 female, 15 male), and 49875 spoofed signals, generated using the same five algorithms that take part in the training set.

4.1.3. Copy-synthesis

Trying to get a more universal model, the same technique as in (Sanchez et al., 2014) is used with both the train and development sets: The human signals are copy-synthesized (at the original 16kHz sampling

[†] http://www.festvox.org
[‡] http://hts.sp.nitech.ac.jp

frequency) using three state-of-the-art vocoders that are widely used in statistical speech processing technologies: AHOCODER (Erro et al., 2014), STRAIGHT (Kawahara et al., 1999) and MLSA (Yoshimura et al., 1999). These additional three signal sets of vocoded “impostors” are used for synthetic model training in some of the experiments, as described in section 5.

4.1.4. Evaluation data

In the evaluation dataset genuine and spoofed signals are included, getting a total of 184000 signals with the same recording conditions as those from the other sets. In this case, 10 different algorithms are used to generate de spoofed signals: the same five that were used for the other sets and 5 different ones, intentionally selected to test the generalization capability of the tested SSD system to face previously unseen attacks. Among these unknown algorithms there is a system (S10) that uses unit selection synthesis, a completely different technology (not based on vocoder parameterization and statistical modeling). This technology was intentionally set aside for the model training material as it can hardly be used for spoofing purposes due to the big amount of signals of the target speaker required to create a quality voice.

4.2. The Blizzard 2012 Database

In order to test the SSD systems with signals completely unrelated with the training material, it was necessary to obtain a representative number of state-of-the-art TTS systems. The Blizzard Challenge (King, 2014) was an interesting choice.

In the field of TTS system design, The Blizzard Challenge is the most popular international event for evaluations. All participants must use a common speech corpus to build a synthetic voice using their TTS systems. Then, some samples of this voice are submitted, so that they can be used in a common subjective evaluation, performed by a large pool of listeners. Undoubtedly, the TTS systems presented to the Blizzard Challenge are a representative sample of the state-of-the-art technology in speech synthesis.

Every year, the Blizzard Challenge organizers distribute the listening evaluation: a set of human recordings and their counterparts synthesized by means of every TTS system that takes part. Since both human and synthetic signals are available, this database can be a good test field for SSD systems.

A wide sample of TTS technologies is present at the Blizzard Challenge: the main groups are statistical or HMM based synthesizers, unit selection based systems and hybrid systems. This last type includes systems that, even using unit selection techniques to generate the speech signal, make use of statistical models in the unit selection process.

In the experiments referred in this paper, we have used the listening evaluation data of the 2012 Blizzard Challenge (King and Karaiskos, 2012). It consists of 11 signal sets, each one with 209 utterances in US English. The set designate A contains the reference human signals, and the system named B is not a participant but a standard unit-selection-based benchmark system. Among the others, we can find statistical systems like E, H and K, unit selections systems like F, G and I, hybrid like C and D and a diphone concatenation system, J.

5. Experiments and Results

We have evaluated the phase-based SSD systems in two experiments using two evaluation sets, as explained in the previous section. For both of them, the systems have been trained with the training and development sets of the ASVspoof DB, including additional signals generated by copy-synthesis of the human subset, using the three vocoders explained in 4.1.3. In the first experiment, the test material belongs to the same database as the training material (the ASVspoof DB) whereas in the second, a completely unrelated evaluation set is used, in order to test the ability of the SSD systems facing completely unknown impostors.

5.1. Evaluation with the ASVSpoof Database

In this experiment the models trained with the training and development material of the ASVspoof DB have been tested with the evaluation part of that database. While the human model has remained the same in the entire experiment, three different training strategies for the synthetic models have been tested (see Table 1):

- M1: Synthetic model developed with the synthetic material provided in the training and development set of the database.
- M2: Synthetic model developed with newly generated synthetic material by copy-synthesis of the human set using three different vocoders: AHOCODER, STRAIGHT and MLSA.
- M3: Synthetic model developed combining the material from M1 and M2.

As mentioned before, the evaluation set is composed of human signals and spoofing signals generated with 10 algorithms, 5 of which are included in the training material. The other five are “unknown”: 4 of them are VC systems with STRAIGHT as vocoder. The 10th system is a unit selection based synthesizer and it is out of the scope of the systems trained in this work.

The results of this experiment are shown in tables 2 and 3. In order to avoid the bias in the average EERs due to the unsurprisingly bad performance of the unit selection based system (S10), we show averaged values for the rest of the systems (*-S10 columns). We also assume that the S10 values are excluded from the analysis in the following lines.

The MFCC baseline gets some decent results with EER around 2%, showing that the classification task is not very demanding for this database.

It is also remarkable that there is no significant difference between known and unknown systems (always excluding the 10th system based in unit selection). For all the systems the performance falls around 10-20% from known to unknown systems when trained with spoofing impostor samples (M1). But actually, with M2 training set (where all the systems are unknown) the performance falls more than for other training sets. That is to say: the slight performance difference cannot be attributed to prior knowledge of the attack method but to other features of particular spoofing systems included in the “unknown” set that affect the performance. This is corroborated by the detailed results of table 3, where it can be seen most of the unknown systems are actually better detected than the known ones. Only the EER for the S6 system, which is particularly bad (for every training set, thus not depending on being known or unknown) makes the average ratio of the unknown systems worse.

The results for the MGD system show a good performance for M1 and M3 training materials, but it degrades for M2 training set. MGD parameters seem to be more affected by the distortions introduced by the statistical modeling process required by real spoofing algorithms, which are not present in vocoded signals used in M2.

RPS based systems get consistently good results in all the training sets and attacking algorithms as can be seen in Table 3, with values well below 1% EER for all the training sets.

Regarding the effect of the different training strategies with RPS parameters, using the attack samples to train the synthetic model of the classifier (M1) performs better than the other strategies. M2, using vocoded material to train the models, produces a poorer but still decent performance, in the same magnitude order than the other strategies. The hypothetical benefit of M2 strategy being capable of producing better results for unknown systems is not shown in the results. Unfortunately, provided the above-mentioned small differentiation between the known and unknown systems, we cannot state whether this generalization feature is or is not true.

M3 strategy, with models created by combination of attack samples and copy-synthesized samples, gets good performance, very close to the M1. Actually, both classifiers are not statistically significant according to the McNemar test (p=0.41). It gives small improvements for some of the “unknown” systems (S8 and S9).

Table 2. Training and evaluation subsets for the different strategies for model training.

	Training		Evaluation	
Model	Human signals	Synthetic signals	Human signals	Synthetic signals
M1	7247	62500	9404	184000
M2	7247	21741 (7247x3)	9404	184000
M3	7247	84241	9404	184000

Table 3. EER in percentage of the different system types tested against the ASVProof database.

SSD System	Known Systems	Unknown Systems-S10	Unknown Systems	All-S10	All
MFCC M1	1.8815	2.1070	9.0998	1.9817	5.4907
MFCC M2	8.9816	11.6447	18.0683	10.1652	13.5250
MFCC M3	1.9262	2.8537	10.0104	2.3384	5.9683
MGD M1	0.9270	1.3086	8.9103	1.0966	4.9187
MGD M2	9.0414	7.4304	14.2777	8.3254	11.6596
MGD M3	2.4529	2.7788	10.2569	2.5977	6.3549
RPS M1	0.1274	0.1562	8.8185	0.1402	4.4730
RPS M2	0.5294	0.6901	10.0970	0.6008	5.3132
RPS M3	0.1361	0.1669	9.1261	0.1498	4.6311

Table 4. EER in percentage of the different systems tested against the ASVProof database.

SSD System	S1	S2	S3	S4	S5	S6	S7	S8	S9	S10
MFCC M1	0.1102	8.4556	0.0360	0.0438	0.7621	1.5449	3.4747	2.4740	0.9343	37.0711
MFCC M2	5.0327	13.1417	8.9981	8.9981	8.7375	11.2048	16.3206	9.4169	9.6362	43.7628
MFCC M3	0.5471	7.6091	0.0690	0.1019	1.3040	2.1734	4.3412	3.1157	1.7846	38.6373
MGD M1	0.1866	1.8559	0.2446	0.2849	2.0629	2.4006	0.7891	1.3186	0.7260	39.3174
MGD M2	1.4017	4.7303	7.3400	6.9395	24.7957	19.2032	2.8205	3.3713	4.3265	41.6672
MGD M3	0.4113	2.5331	0.9455	0.9500	7.4247	6.5415	1.0769	1.9259	1.5708	40.1692
RPS M1	0.2661	0.1695	0.0217	0.0360	0.1439	0.5147	0.0080	0.0912	0.0108	43.4680
RPS M2	1.0478	0.7359	0.1625	0.1437	0.5571	2.0239	0.1286	0.2834	0.3243	47.7249
RPS M3	0.2814	0.1770	0.0152	0.0363	0.1707	0.5711	0.0108	0.0755	0.0101	44.9628

5.2. Using the Blizzard 2012 Database

The second experiment aims to analyze the performance of the SSD systems when they are confronted with completely unrelated signals, both natural and spoofed. Besides the unknown spoofing algorithm used, these signals would be acquired in a completely different channel, and thus the intrinsic robustness of the different SSDs to the channel-mismatch issue will also be evaluated.

As mentioned before, we will use the Blizzard 2012 Database with 10 voice adapted TTS (B-K) plus the natural voice. 3 of the TTS in this Challenge (E, H and K) are statistical synthesizers which use HMM based models of certain speech parameters which, in the synthesis phase, will feed a vocoder to produce the speech signal.

The rest of the systems use unit selection or hybrid technologies for synthesis, which means that they concatenate segments of natural signals and therefore do not use any vocoder. As was the case with S10 algorithm in the previous experiment, these systems are out of the scope of the SSDs evaluated here, and should been addressed specifically in future works. On the other hand, unit selection based technology might not be suitable for spoofing in some applications which require live conversation (call-center applications, for instance), as unit selection technology requires a relative large speech database to produce natural speech, and can be easily detected by human ears (Wester et al., 2015).

In this experiment all the SSDs and the three training strategies have been evaluated, and the results are shown in table 5. The EER of the system is obtained by testing every synthetic subset against the human counterpart.

Table 5. EER in percentage of the different systems tested against the ASVSpoof database.

SSD System	B	C	D	E	F	G	H	I	J	K
MFCC M1	38.2775	44.0191	17.2249	0.0000	27.7512	28.7081	3.8278	22.0096	62.2010	0.0000
MFCC M2	48.8038	48.8038	52.6316	0.9569	29.1866	35.4067	18.1818	24.8804	14.8325	2.8708
MFCC M3	43.5407	45.9330	20.5742	0.0000	25.3589	29.1866	3.8278	22.0096	10.0478	0.0000
MGD M1	59.8086	73.2057	8.6124	2.3923	78.4689	79.4258	7.6555	63.1579	42.5837	5.2632
MGD M2	62.6794	23.4450	8.1340	5.7416	44.4976	32.5359	3.3493	17.7033	23.4450	3.3493
MGD M3	59.8086	37.7990	5.2632	3.8278	60.7656	54.0670	3.8278	33.0144	27.2727	3.8278
RPS M1	49.2823	69.3780	40.6699	0.0000	69.3780	2.8708	0.0000	32.0574	72.2488	0.0000
RPS M2	34.9282	58.3732	11.0048	0.0000	66.0287	2.8708	0.0000	19.1388	6.2201	0.0000
RPS M3	40.6699	61.7225	15.7895	0.0000	63.1579	3.8278	0.0000	23.9234	14.3541	0.0000

The first evident result is that none of the systems or the strategies is able to correctly detect the unit selection based systems. This is consistent with the results of the S10 system in the previous database. The error level is comparable in both experiments, which means that it is not due to the signals being unknown or to the channel-mismatch, but it reflects the intrinsic inability to detect unit selection systems without such samples in the model training.

Regarding the vocoder based synthesis systems the results depend on the system. The baseline SSD gets good results for some TTS but its performance depends upon the training strategy and system. The MGD based system seems to be sensitive to the channel mismatch problem, because the detection rate is not so good. The RPS based system, by the contrary, obtains consistent error-free classification regardless the TTS or the training strategy, suggesting robustness to the channel and attacking system variation. The average errors for vocoder based and unit selection based systems is shown in table 6.

Table 6. Average EER in percentage for the different types of synthetic signals.

SSD System	Average vocoder based systems (E,H,K)	Average unit selection & hybrid systems	Average All
MFCC M1	1.2759	34.3131	24.4019
MFCC M2	7.3365	36.3636	27.6555
MFCC M3	1.2759	28.0930	20.0478
MGD M1	5.1037	57.8947	42.0574
MGD M2	4.1467	30.3486	22.4880
MGD M3	3.8278	39.7129	28.9474
RPS M1	0.0000	47.9836	33.5885
RPS M2	0.0000	28.3664	19.8565
RPS M3	0.0000	31.9207	22.3445

Regarding the training strategy, the experiment shows diverse behaviours depending on the SSD and the type of impostors. For the baseline system the M3 strategy, combining training samples from spoofing signals and vocoded ones seems the best approach. Conversely, for the MGD and the RPS systems the M2 strategy (with just vocoded signals) seems to be the best (attending to the average EER for all the systems). This result is mainly due to performance with the unit selection systems, which, although very bad in all the cases, is better for the M2 training strategy. It can be hypothesised that models trained with attacking samples (M1) are too specific to capture the features of other synthesis techniques, while vocoded signals, being of higher quality and closer to the natural signals produce more general models better suited to detect unknown signals.

Considering all the results, there is one interesting observation worth noting: RPS which uses purely phase information achieves similar performance by using different training conditions. On the other hand, MFCC or MGD which also consider magnitude information varies a lot. It means that the magnitude spectrum is distorted a lot after the modelling process by speech synthesis or voice conversion what does not happen for the vocoded speech, which produces very high quality synthetic signals. It is possible that MGD and MFCC systems are actually modelling the distortions in the magnitude spectra produced by the speech processing technology.

6. Conclusions

In this paper we have reviewed two phase based methods to detect spoofing using synthetic speech: both are based in GMM models for natural and synthetic signals but one of them uses Modified Group Delay parameters to train the models while the other uses DCT-mel-RPS parameters. We also use a MFCC based system as baseline. We have focused on attacks created with speaker adapted synthetic speech and voice conversion systems which use parameter manipulation followed by speech generation using vocoders, as they are the most feasible methods to generate the spoofing signals.

We have evaluated these systems using two databases, with training material coming only from one of them in order to evaluate the systems with completely unrelated signals (including acquisition channel). This evaluation intends to simulate real application scenarios and to assess the generalization abilities of these countermeasures.

We have also evaluated different training strategies, aiming to address the problem of obtaining suitable training data for the spoofing signal model. Hence, we have developed models from "real" spoofing signals but also with copy-synthesizes signals using three of the most common vocoders used in current adapted synthetic speech and voice conversion systems.

The results show that the systems can achieve a good performance, which is maintained even with completely unrelated signals coming from other database. Both phase-based systems improve the baseline results. The best training strategy appears to be using spoofing samples, but adding vocoded signals can help improving results with unknown signals. For the RPS based classifier using both types of signals to train the model has no significant downside. More extensive evaluation is needed with different attacking technologies and signals to definitely asses the convenience of such training strategies.

Although they are not the target of the SSDs developed in this work, we have kept the unit selection systems in the test material. As it was expected, the SSD systems trained with vocoder based synthetic signals do not work with unit selection based ones. It has to be studied if phase based systems like the ones here presented, trained with appropriate signals, and could model this kind of synthetic impostors.

Acknowledgements

This work has been partially supported by the Basque Government (Ber2Tek Project, IE12-333) and the Spanish Ministry of Economy and Competitiveness (SpeechTech4All project, TEC2012-38939-C03-03).

References

Alegre, F., Amehraye, A., Evans, N., 2013a. Spoofing countermeasures to protect automatic speaker verification from voice conversion, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 3068–3072. doi:10.1109/ICASSP.2013.6638222

Alegre, F., Vipperla, R., Amehraye, A., Evans, N., 2013b. A new speaker verification spoofing countermeasure based on local binary patterns, in: INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association.

Alsteris, L.D., Paliwal, K.K., 2007. Short-time phase spectrum in speech processing: A review and some experimental results. Digit. Signal Process. 17, 578–616. doi:10.1016/j.dsp.2006.06.007

Black, A.W., Tokuda, K., 2005. The Blizzard Challenge 2005: Evaluating corpus-based speech synthesis on common datasets, in: INTERSPEECH 2005 - Eurospeech, 9th European Conference on Speech Communication and Technology, Lisbon, Portugal, September 4-8, 2005. pp. 77–80.

De Leon, P.L., Hernaez, I., Saratxaga, I., Pucher, M., Yamagishi, J., 2011. Detection of Synthetic Speech for the Problem of Imposture. 2011 IEEE Int. Conf. Acoust. Speech Signal Process. ICASSP 4844–4847.

De Leon, P.L., Pucher, M., Yamagishi, J., Hernaez, I., Saratxaga, I., 2012. Evaluation of Speaker Verification Security and Detection of HMM-Based Synthetic Speech. IEEE Trans. Audio. Speech. Lang. Processing 20, 2280–2290. doi:10.1109/TASL.2012.2201472

Erro, D., Sainz, I., Navas, E., Hernaez, I., 2014. Harmonics Plus Noise Model Based Vocoder for Statistical Parametric Speech Synthesis. IEEE J. Sel. Top. Signal Process. 8, 184–194. doi:10.1109/JSTSP.2013.2283471

Hegde, R.M., Murthy, H.A., Gadde, V.R.R., 2007. Significance of the Modified Group Delay Feature in Speech Recognition. IEEE Trans. Audio, Speech Lang. Process. 15, 190–202. doi:10.1109/TASL.2006.876858

Jin, Q., Toth, A.R., Black, A.W., Schultz, T., 2008. Is voice transformation a threat to speaker identification?, in: ICASSP. pp. 4845–4848.

Kawahara, H., Masuda-Katsuse, I., de Cheveigné, A., 1999. Restructuring speech representations using a pitch-adaptive time–frequency smoothing and an instantaneous-frequency-based F0 extraction: Possible role of a repetitive structure in sounds. Speech Commun. 27, 187–207. doi:10.1016/S0167-6393(98)00085-5

King, S., 2014. Measuring a decade of progress in Text-to-Speech. Loquens 1, e006. doi:10.3989/loquens.2014.006

King, S., Karaiskos, V., 2012. The Blizzard Challenge 2012, in: Proc. of The Blizzard Challenge 2012.

Kinnunen, T., Wu, Z., Lee, K.A., Sedlak, F., Chng, E.S., Li, H., 2012. Vulnerability of speaker verification systems against voice conversion spoofing attacks: The case of telephone speech, in: 2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, pp. 4401–4404. doi:10.1109/ICASSP.2012.6288895

Masuko, T., Tokuda, K., Kobayashi, T., 2000. IMPOSTURE USING SYNTHETIC SPEECH AGAINST SPEAKER VERIFICATION BASED ON VERIFICATION BASED ON, in: ICSLP.

Pellom, B.L., Hansen, J.H.L., 1999. An experimental study of speaker verification sensitivity to computer voice-altered imposters. 1999 IEEE Int. Conf. Acoust. Speech, Signal Process. Proceedings. ICASSP99 (Cat. No.99CH36258) 2, 837–840. doi:10.1109/ICASSP.1999.759801

Rosenberg, A.E., 1976. Automatic speaker verification: A review. Proc. IEEE 64, 475–487. doi:10.1109/PROC.1976.10156

Sanchez, J., Saratxaga, I., Hernaez, I., Navas, E., Erro, D., 2014. A Cross-vocoder Study of Speaker Independent Synthetic Speech Detection using Phase Information, in: Interspeech. Singapore, pp. 1663–1667.

Sanchez, J., Saratxaga, I., Hernaez, I., Navas, E., Erro, D., Raitio, T., 2015. Towards a Universal Synthetic Speech Spoofing Detection using Phase Information. IEEE Trans. Inf. Forensics Secur. PP, 1–1. doi:10.1109/TIFS.2015.2398812

Saratxaga, I., Hernaez, I., Erro, D., Navas, E., Sanchez, J., 2009. Simple representation of signal phase for harmonic speech models. Electron. Lett. 45, 381. doi:10.1049/el.2009.3328

Saratxaga, I., Hernáez, I., Odriozola, I., Navas, E., Luengo, I., Erro, D., 2010. Using Harmonic Phase Information to Improve ASR Rate, in: Interspeech. pp. 1185–1188.

Saratxaga, I., Hernaez, I., Pucher, M., Navas, E., Sainz, I., 2012. Perceptual Importance of the Phase Related Information in Speech, in: Interspeech. ISCA, Portland, OR, pp. 1448–1451. doi:10.1.1.396.4789

Satoh, T., Masuko, T., Kobayashi, T., Tokuda, K., 2001. A robust speaker verification system against imposture using an HMM-based speech synthesis system., in: Dalsgaard, P., Lindberg, B., Benner, H., Tan, Z.-H. (Eds.), Interspeech. ISCA, pp. 759–762.

Sizov, A., Khoury, E., Kinnunen, T., Wu, Z., Marcel, S., 2015. Joint Speaker Verification and Antispoofing in the i-Vector Space. IEEE Trans. Inf. Forensics Secur. 10, 821–832. doi:10.1109/TIFS.2015.2407362

Tokuda, K., Zen, H.Z.H., Black, A.W., 2002. An HMM-based speech synthesis system applied to English. Proc. 2002 IEEE Work. Speech Synth. 2002. 2–5. doi:10.1109/WSS.2002.1224415

Wester, M., Wu, Z., Yamagishi, J., 2015. Human vs Machine Spoofing Detection on Wideband and Narrowband Data, in: Interspeech 2015. Singapore, pp. 1–5.

Wu, Z., Chng, E.S., Li, H., 2012. Detecting Converted Speech and Natural Speech for anti-Spoofing Attack in Speaker Recognition. Interspeech 2–5.

Wu, Z., Evans, N., Kinnunen, T., Yamagishi, J., Alegre, F., Li, H., 2015a. Spoofing and countermeasures for speaker verification: A survey. Speech Commun. 66, 130–153. doi:10.1016/j.specom.2014.10.005

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., 2014. ASVspoof 2015 : Automatic Speaker Verification Spoofing and Countermeasures Challenge Evaluation Plan [WWW Document]. URL http://www.spoofingchallenge.org/asvSpoof.pdf

Wu, Z., Kinnunen, T., Evans, N., Yamagishi, J., Hanilci, C., Sahidullah, M., Sizov, A., 2015b. ASVspoof 2015: the First Automatic Speaker Verification Spoofing and Countermeasures Challenge, in: Poc. Interspeech 2015.

Wu, Z., Xiao, X., Chng, E.S., Li, H., 2013. Synthetic speech detection using temporal modulation feature, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, pp. 7234–7238. doi:10.1109/ICASSP.2013.6639067

Yamagishi, J., Nose, T., Zen, H., Ling, Z.-H., Toda, T., Tokuda, K., King, S., Renals, S., 2009. Robust Speaker-Adaptive HMM-Based Text-to-Speech Synthesis. IEEE Trans. Audio. Speech. Lang. Processing 17, 1208–1230. doi:10.1109/TASL.2009.2016394

Yamagishi, J., Usabaev, B., King, S., Watts, O., Dines, J., Oura, K., Tokuda, K., Karhila, R., Kurimo, M., 2010. Thousands of Voices for HMM-Based Speech Synthesis–Analysis and Application of TTS Systems Built on Various ASR Corpora. IEEE Trans. Audio, Speech Lang. Process. 18, 984–1004. doi:10.1109/TASL.2010.2045237

Yegnanarayana, B., Murthy, H., 1992. Significance of group delay functions in spectrum estimation. Signal Process. IEEE … 40, 2281–2289. doi:10.1109/78.157227

Yoshimura, T., Tokuda, K., Masuko, T., Kobayashi, T., Kitamura, T., 1999. Simultaneous modeling of spectrum, pitch and duration in HMM-based speech synthesis. Proc. Eurospeech 2347–2350.

Zhu, D., Paliwal, K.K., 2004. Product of power spectrum and group delay function for speech recognition, in: 2004 IEEE International Conference on Acoustics, Speech, and Signal Processing. IEEE, pp. I–125–8. doi:10.1109/ICASSP.2004.1325938